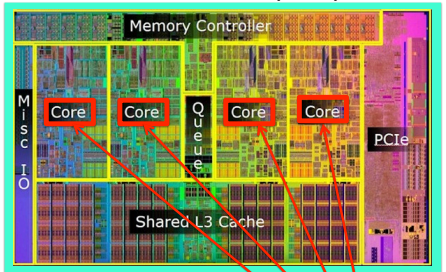
Graphics Processor Units (GPUs) Explained

Mark Govett
NOAA Earth System Research Laboratory

CPU: Nahalem (2009)



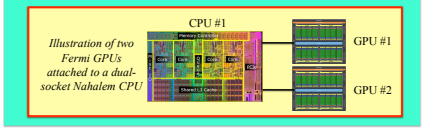
1 Processing Unit

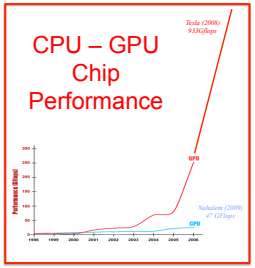
Computational Capability

- **4 Intel I7 cores**
- 4 cores execute instructions simultaneously
- **focus on single-thread performance**
  - speculative execution
  - 32K L1, 256K L2 cache
- 8 MB Shared L3 cache

The Basics

- GPUs are a CPU co-processor
- Developed for the video gaming industry
- Millions of GPUs are sold every year
- GPUs come standard with many desktop and laptop systems





CPU – GPU Comparison at a Glance

CHIP TYPE	CPU Nahalem	GPU NVIDIA Tesla	GPU NVIDIA Fermi
Cores	4	240	512
Parallelism	Medium Grain	Fine Grain	Fine Grain
Performance Single Precision Double Precision	47 GFlops	933 GFlops 60 GFlops	1040 GFlops 500 GFlops
Power Consumption	130W	150W	220W
Transistors	730 million	1.4 bilion	3.0 billion

Why are GPUs So Fast?

- Design maximizes computational efficiency
- **Chip space is dominated by processing units**
- High number of compute cores
- Cores are simple, lightweight, low-power

GPU: NVIDIA Tesla (2010)

32 Processing Units

Computational Capability

- **512 cores executing simultaneously**
- 16 Streaming Processors (SP)
  - 32 cores (a *warp*) execute the same instruction simultaneously
  - dual issue warp scheduling
  - rapid context switching

Getting to Operational PetaFlop Computing

Operational Computing at NCEP

Where We Are Today

~180 TeraFlops (2 systems)
5000 IBM Power 6
68th, 69th fastest on Top500 (Nov2009)

Power
0.5 MegaWatts

Reliability
Use two 90 TeraFlop Systems

99.9 reliability requirement

NCEP Operational Computing Facility

Research Computing: DOE Jaguar

State of the Art in CPU Computing Innovative building design, cooling, power efficiency

2.3 PetaFlops
250,000 AMD cores
284 cabinets of computing

Power
7-10 MegaWatts
(sufficient for 8-10,000 homes)

Reliability
MTBF: estimated in **hours**

Cost
Facilities ($73M), System (~ $100M)
Annual Power (~ $4M)

DOE Oak Ridge Computing Facility

Valmont Power Plant ~225 MegaWatts Boulder, CO

GPU Cluster Computing

Alternative Fermi System

1.0 PetaFlops
1000 NVIDIA Fermi GPUs
500 Intel CPU Nodes
10 cabinets of computing

Power
0.5 MegaWatts

Reliability
MTBF: Estimated in **weeks**

Cost
System (~ $5M)
Annual Power (~ $250K)

NOAA Boulder Computing Facility

Alternative Computing Technologies

ATI Radeon GPU (2010)
5.0 TeraFlops Performance
Graphics card only
No HPC language support

Intel SandyBridge CPU (2011)
8 cores + "GPU" extensions